

MULTILINEAR SUBSPACE CLUSTERING

Eric Kernfeld*, Nathan Majumder†, Shuchin Aeron†, and Misha Kilmer†

* University of Washington, Seattle, WA, USA. † Tufts University, Medford, MA, USA.

ABSTRACT

In this paper we present a new model and an algorithm for unsupervised clustering of 2-D data such as images. We assume that the data comes from a union of multilinear subspaces (UOMS) model, which is a specific structured case of the much studied union of subspaces (UOS) model. For segmentation under this model, we develop Multilinear Subspace Clustering (MSC) algorithm and evaluate its performance on the YaleB and Olivetti image data sets. We show that MSC is highly competitive with existing algorithms employing the UOS model in terms of clustering performance while enjoying improvement in computational complexity.

Index Terms – subspace clustering, multilinear algebra, spectral clustering

1. INTRODUCTION

Most clustering algorithms seek to detect disjoint clouds of data. However, in high-dimensional statistics, data can become very sparse, and these types of methods have trouble dealing with noise. In fact, a completely new approach to the geometry of clustering has recently made headway in the analysis of high-dimensional data sets. Called subspace clustering, this approach assumes that data come from subspaces offset at angles, rather than from clouds offset by gaps, the so called Union of Subspaces (UOS) model [1, 2, 3]. Applications have included detection of tightly correlated gene clusters in genomics [4], patient-specific seizure detection from EEG data [5], and image segmentation [6].

All subspace clustering methods must embed data in \mathbb{R}^n . However, in some of the high-dimensional data sets where subspace clustering has been applied, the initial structure of the data is not a vector but rather a matrix or tensor (multiway array). Examples include the auditory temporal modulation features in [7], the image patches in [6], and raw EEG data under the “sliding-window approach” [5]. We seek to develop a clustering method that incorporates the geometric innovation of subspace clustering without vectorizing these higher-order arrays. To do this, we formulate an algebraic generative model for the data, along with methods for inference.

The Subspace Clustering Problem and a Multilinear Variant - Mathematically, the subspace clustering problem is described as follows. Given a set of points x_n , $n = 1 \dots N$, suppose each point is an element of one of the K subspaces. The problem is to decide membership for each of the N points. For simplicity, we treat K as known.

In order to take advantage of patterns in two-way data, we modify the assumptions of the subspace clustering problem. Rather than modeling the data as a union of subspaces, we assume they come from a union of *tensor products* of subspaces [8]. Given subspaces $\mathcal{U} \subset \mathbb{R}^n$ and $\mathcal{V} \subset \mathbb{R}^m$, suppose the columns of \mathbf{U} form a basis of \mathcal{U} and likewise for \mathbf{V} and \mathcal{V} . The tensor product $\mathcal{U} \otimes \mathcal{V}$ is the set $\{\mathbf{A} | \mathbf{A} = \mathbf{U}\mathbf{Y}\mathbf{V}^T\}$, where \mathbf{Y} is a $\dim(\mathcal{U}) \times \dim(\mathcal{V})$ matrix. In other words, this is a set of matrices with (column/row) space confined to $(\mathcal{U}/\mathcal{V})$. We refer to this model as the union of multilinear subspaces (UOMS) model and we call this the multilinear subspace clustering (MSC) problem. **Note that while $\mathcal{U} \otimes \mathcal{V}$ is a tensor-subspace of the tensor space $\mathbb{R}^n \otimes \mathbb{R}^m$, not all subspaces of the tensor space is a tensor subspace [8]. Therefore we are assuming a tensor-subspace structure on the clusters under the UOMS model.**

The difference between the generative models for UOS and UOMS is clarified in Algorithms 1, 2.

Algorithm 1 UOS Data Generation: N points, K clusters of latent dimension d and ambient dimension D

Given $\{\mathbf{U}_1, \dots, \mathbf{U}_K\} \in \mathbb{R}^{D \times d}$,

Repeat N times:

Draw k from $\{1, \dots, K\}$

Draw a random length- d vector y_n

Compute datum $\mathbf{x}_n = \mathbf{U}_k \mathbf{y}_n$

Algorithm 2 UOMS Data Generation: N points, K clusters of latent dimension $d_u d_v$ and ambient dimension $D_u D_v$

Given $\{\mathbf{U}_1, \dots, \mathbf{U}_K\} \in \mathbb{R}^{D_u \times d_u}$, $\{\mathbf{V}_1, \dots, \mathbf{V}_K\} \in \mathbb{R}^{D_v \times d_v}$

Repeat N times:

Draw k from $\{1, \dots, K\}$

Draw a random d_u by d_v matrix \mathbf{Y}_n

Compute datum $\mathbf{A}_n = \mathbf{U}_k \mathbf{Y}_n \mathbf{V}_k^T$

Relation of UOMS to existing subspace models - Note that the UOS model with single subspace (one cluster) is related to the Principal Component Analysis (PCA). Similarly

This work is part of the first author’s senior honors thesis at Tufts University. Nathan Majumder, Shuchin Aeron and Misha Kilmer were supported by NSF grant 1319653.

the UOMS with one cluster is closely related to separable covariance models [9] and also 2D-PCA[10]. Further in [9], extensions of this idea to 3-D, 4-D,... data is shown to be equivalent to HOSVD and Tucker decompositions [11] that have been useful for dimensionality reduction for image ensembles [12]. Further such multilinear subspace models have been used in machine learning [13, 14]. In this paper we study an extension of these models by considering a *union* of such structured subspaces.

Algorithm 3 Thresholded Subspace Clustering (TSC)

Input

$\mathbf{X} \in \mathbb{R}^{D \times N}$ holding data vectors \mathbf{x}_i , number of clusters K , threshold ($1 \leq q \leq N$)

Procedure

Normalize \mathbf{x}_i , Compute adjacency matrix $\mathbf{C} = |\mathbf{X}^\top \mathbf{X}|$
Set all but q highest values of each row of \mathbf{C} to zero
 $\mathbf{C}_{i,j} = \exp(-2 * \cos^{-1}(\mathbf{C}_{ij}))$
Perform normalized spectral clustering on \mathbf{C}

Output

A vector in \mathbb{R}^N with clustering labels for all \mathbf{x}_i

Algorithm 4 Sparse Subspace Clustering (SSC)

Input

$\mathbf{X} \in \mathbb{R}^{D \times N}$ holding data vectors \mathbf{x}_i , $\lambda_1, \lambda_2 \geq 0$
the number of clusters K

Procedure

Solve $\arg \min_{\mathbf{C}, \mathbf{E}} \|\mathbf{X} - \mathbf{X}\mathbf{C} - \mathbf{E}\|_F^2 + \lambda_1 \|\mathbf{C}\|_1 + \lambda_2 \|\mathbf{E}\|_1$
s.t. $\mathbf{C}_{ii} = 0$
Form $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^\top$
Apply spectral clustering on \mathbf{W}

Output

A vector in \mathbb{R}^N with clustering labels for all \mathbf{x}_i

Clustering under the UOS model - There are many algorithms exploiting the UOS model for clustering, but we focus on two general methods that form an affinity matrix among data points followed by spectral clustering [15]. The first, called Thresholded Subspace Clustering (TSC), is introduced in [2]. This provably reliable and robust subspace clustering algorithm involves constructing a weighted graph where nodes represent the data points and edges represent the connectivity of any two points. The inner product between the data points is used as the edge weight with the idea that points in the same subspace will generally have a higher inner product than points in different subspaces. The symmetric adjacency matrix for the graph is then thresholded, setting all but the q highest weights in each row to zero, in order to filter out noise. The second method, called Sparse Subspace Clustering (SSC) [16] involves expressing each point as a linear combination of the other points in the dataset. The algorithm finds the sparsest possible linear representation of each data point in terms of other data points – achieved by minimizing the ℓ_1 -norm – with the idea that the points used will come from the

same subspace as the point in question. A weighted graph is then formed with an adjacency matrix found using the sparse representations of each point. Both the TSC and SSC algorithms, taken from [2] and [16] respectively, are detailed in Algorithms 3 and 4.

2. CLUSTERING UNDER THE UOMS MODEL

In the case of two-way data, our data points would be a collection of N matrices $\mathbf{A}_n \in \mathbb{R}^{D_U \times D_V}$ such that the columns come from a union of subspaces $\mathcal{U}_1 \cup \dots \cup \mathcal{U}_K$ and the rows come from a union of subspaces $\mathcal{V}_1 \cup \dots \cup \mathcal{V}_K$. To take advantage of this fact and find these \mathcal{U}_i and \mathcal{V}_i subspaces, one method would be to cluster all $D_V N$ columns and all $D_U N$ rows separately; however, this is an expensive solution. Instead, we randomly select a single column and a single row from each matrix and cluster these. We stack the random columns side by side to form a $D_U \times N$ matrix \mathbf{X}_{cols} and transpose and stack our random rows side by side to form a $D_V \times N$ matrix \mathbf{X}_{rows} . The i th column of each of these matrices comes from the i th ($i = 1, \dots, N$) data matrix \mathbf{A}_i . We then perform a clustering algorithm on \mathbf{X}_{rows} and \mathbf{X}_{cols} separately, but pause after obtaining the symmetric adjacency matrix \mathbf{C} in each case. We repeat this process for T trials, ending up with $2T$ adjacency matrices, which we then combine in one of a few possible ways.

Possible combination methods are detailed subsequently. Combining these adjacency matrices can be thought of as condensing multiple graph realizations to obtain a single weighted graph representing our clustering problem. Once we have our condensed graph, we perform spectral clustering on it to achieve the segmentation of our original points. Algorithm 5 outlines the steps described above.

Algorithm 5 Multilinear Subspace Clustering (MSC)

Input:

Data $\mathbf{A}_1, \dots, \mathbf{A}_N \in \mathbb{R}^{D_c \times D_r}$, number of clusters K
Clustering Method (TSC or SSC), number of trials T

Procedure:
For T trials

form $\mathbf{X}_{cols} \in \mathbb{R}^{D_c \times N}$ with i -th column randomly selected from \mathbf{A}_i

form $\mathbf{X}_{rows} \in \mathbb{R}^{D_r \times N}$ with i -th column randomly selected from \mathbf{A}_i^\top

Run clustering (TSC or SSC) on \mathbf{X}_{cols} and \mathbf{X}_{rows} to get $\mathbf{C}_{cols} \in \mathbb{R}^{N \times N}$ and $\mathbf{C}_{rows} \in \mathbb{R}^{N \times N}$

End For

Combine all \mathbf{C}_{cols} and \mathbf{C}_{rows} into a single adjacency matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$

Perform spectral clustering on \mathbf{C}

Output:

A vector in \mathbb{R}^N with clustering labels for all \mathbf{A}_i

Combining the Graph Realizations - We now discuss several (heuristic) methods for combining the adjacency matrices obtained at each iteration of MSC.

1. **Addition:** One simple method is to add the $2T$ adjacency matrices together.
2. **Thresholding:** Add the matrices together followed by thresholding, setting all but the q highest edges per row to zero. A possible choice of threshold for this method would be the average number of data points per cluster – if this number is known – minus one (to count out the point itself).
3. **Filtering by Quantile:** A "quantile" method involves choosing a parameter l and taking the l -th highest weight at each edge out of all the adjacency matrices. The choice of l poses an obstacle, as there is not a given value that will be optimal for all graphs.
4. **Projection:** Project each individual adjacency matrix's columns onto its leading K singular vectors (corresponding to largest singular values) before adding the instances. However, the fact that each matrix is projected onto its leading singular vectors before sharing any information with other graph realizations could lead to loss of quality.

Remark - These methods are by no means exhaustive. In particular, the problem of combining various graph realizations for the same problem instance by itself is an interesting avenue of research.

Algorithmic Complexity - For N data points of dimension D , TSC has algorithmic complexity $\mathcal{O}(DN^2)$. Therefore, if we are comparing TSC on vectorized 2-way data against MSC using TSC on the same data in matrix form, the MSC data points will be matrices of size $D_c \times D_r$ where $D = D_c D_r$. At each iteration of MSC, we form the matrices of size $D_c \times N$ and $D_r \times N$ for the column space and row space respectively. Since we then perform TSC on these matrices, the algorithmic complexity at each iteration will be $\mathcal{O}(D_c N^2 + D_r N^2)$ or approximately $\mathcal{O}(\sqrt{D} N^2)$ when $D_c \approx D_r$. For the projection method, which is the computationally most expensive, when $K \ll N$, using randomized SVD a computational cost of $\mathcal{O}(N^2 \log K)$ [17] is incurred. Therefore, for T iterations of MSC, we have $\mathcal{O}(T \sqrt{D} N^2)$ compared to $\mathcal{O}(DN^2)$ for TSC. Therefore if we can pick a number of trials T such that $T \ll \sqrt{D}$, MSC will be cheaper. This obviously leads to a possible conclusion that MSC will be a better choice for large data while TSC will be more realistic for data of smaller dimensions. The computational complexity of the SSC algorithm is $\mathcal{O}(DN^3)$, which for large D and N becomes more prohibitive compared to TSC as well as MSC.

3. NUMERICAL RESULTS

3.1. Yale Face Database

We first test our Multilinear Subspace Clustering algorithm on images of faces with fixed pose and expression under different illumination conditions from the Yale Face Database

Table 1. The clustering error for our four methods for condensing the graph obtained in MSC.

% MSC Clustering Errors		Number of Clusters		
		2	5	10
Method	addition	12.21	44.16	48.96
	projection	4.06	20.19	38.70
	quantile	20.75	47.37	45.52
	threshold	11.13	35.52	46.15

B. These images are of size 192×168 pixels and we use 64 images each of 37 different people. The parameters that we vary are the number of trials T and the method of combining graph realizations. Over numbers of clusters ranging from 2 to 10, we compare results to those obtained in [2] and [16] using TSC and SSC. We measure success based on a clustering error that measures the fraction of misclassified points and is defined in [2].

3.2. Results of Different Methods for Combining Graphs

We first test the various methods of condensing our graph that we detailed in section 2. We set $T = 15$ as a constant number of trials and find the clustering error for 2, 5, and 10 clusters using all four methods. The results, depicted in Table 1, show that the method of projecting the columns of the adjacency matrices onto their leading eigenvectors before adding them produces the smallest clustering error by far. Therefore we use this method for future results shown in this paper.

3.3. Varying the Number of Trials

Using the project-first method of combining our graph realizations, we next test the effect of varying the number of trials used for MSC. The goal is to find a number independent of the number of data points that is small enough such that MSC remains as efficient as possible (based on the discussion on computational complexity) while minimizing the clustering error. From Table 2 We note that the clustering error begins to plateau at 100 - 200 trials. We are randomly selecting columns and rows from each matrix, so ideally we get a different fiber at each iteration. Setting the number of trials to be more than the number of fibers does not make sense if we do not want to use a fiber more than once. In the following subsection, when we discuss MSC using SSC, we use 100 trials, as it seems to be where increasing the trials begins to have less effect on performance.

3.4. Varying the SSC Parameters

The SSC code taken from [16] requires parameter that were previously optimized for vectorized data. We therefore vary three of these inputs to achieve the best results. The first two parameters allow for the algorithm to reject outliers and allow

Table 2. The MSC using TSC clustering error data

% MSC Clustering Errors		Number of Clusters		
		2	5	10
Number of Trials	10	5.50	22.61	43.85
	20	4.44	19.01	40.73
	50	3.37	15.68	38.28
	100	2.65	14.43	36.82
	200	2.32	13.73	34.95

Table 3. MSC using SSC

% MSC CE			Number of Clusters		
			2	5	10
Constraints	Outliers ON	Affine ON	10.02	58.17	70.26
	Outliers ON	Affine OFF	2.71	21.22	40.68
	Outliers OFF	Affine ON	25.57	59.51	68.80
	Outliers OFF	Affine OFF	25.57	59.51	68.80

for affine subspaces respectively. If the outlier parameter is "true", the SSC code will dismiss outlying entries, and if the affine parameter is "true", the code will broaden its scope to include affine subspaces. The numerical results can be seen in Table 3. Interestingly, the best results come from allowing outlier rejection but disallowing affine subspaces.

Note: This is slightly counterintuitive for MSC, as SSC is being run on a matrix of single fibers, not on an entire data point at each iteration. Therefore, one might think that since many of the fibers will have no connection to each other, random fibers would get thrown out as outliers, when really we would want to get rid of entire outlying data points, not single fibers. This is an important avenue for future research.

Overall Performance Comparisons - We first compare the use of TSC on vectorized data to MSC using TSC at each iteration (on data in matrix form). The average clustering errors for TSC given in [2] are (without preprocessing) 12.42%, 29.17%, 39.84% for 2, 5, and 10 clusters respectively. The full set of data from varying the number of MSC trials can be seen in Table 2, but for 100 trials, the number at which the decrease in error begins to plateau, we obtain average clustering errors of 2.65%, 14.43%, and 36.82% for 2, 5, and 10 clusters respectively on the same data. Lowering the number of trials we use to 50 and even 20, we still obtain better clustering errors using MSC for low numbers of clusters and comparable clustering errors for 10 clusters.

Because each face datum is so large, we cannot compare the SSC algorithm to MSC using SSC. Vectorizing the data yields points of dimension $192 * 168 = 32256$, which is far too large to use the SSC algorithm from [16], given the optimization problem of constructing a sparse representation

Table 4. The clustering errors for each algorithm on the Yale Faces dataset using 2, 5, and 10 clusters

% Clustering Errors		Number of Clusters		
		2	5	10
Algorithm	TSC	12.42	29.17	39.84
	MSC using TSC	2.65	14.43	36.82
	SSC	timed out		
	MSC using SSC	2.31	15.43	36.67

Table 5. The clustering errors for each algorithm on the Olivetti Faces dataset using 2, 5, and 10 clusters

% Clustering Errors		# Clusters		
		2	5	10
Algorithm	TSC	7.70	23.78	25.01
	MSC using TSC	4.40	14.04	20.79
	SSC	2.90	9.72	18.10
	MSC using SSC	2.30	10.70	16.28

for each point. This therefore is an immediate advantage for MSC, as SSC is impractical for any very large data. We do report clustering errors for MSC using SSC at each iteration, as can be seen in the previous subsection. The optimized results are, for 2, 5, and 10 clusters respectively, 2.31%, 15.43%, and 36.67%, which are comparable to those obtained from MSC using TSC on the same data. The results from each algorithm are shown in Table 4.

3.5. Olivetti Face Database

We look at results of the algorithms on the Olivetti Faces data, found at [18]. It includes 400 images of size 64×64 pixels, 10 each of 40 different subjects. For each test, we perform and average results from 100 runs, each of which chooses a random set of subjects and takes all 10 images from each. The results are shown in Table 5.

4. CONCLUSIONS AND FUTURE WORK

In this paper we presented a new model, namely the UOMS model and presented an algorithm unsupervised clustering under this model. We showed that the resulting algorithm is competitive with existing methods while being computationally more efficient. For future work, we will investigate how to deal with the outliers when we are drawing rows and columns from the data and not the entire data itself. Another important avenue is to develop a systematic method with provable guarantees for combining various graph realizations. As is well known, subspace clustering performance depends on the distribution of the data. Therefore, if the relative importance of different instances can be characterized one may perform a weighted graph combination. We will also extend this method for automatic clustering of 3-D data sets such as action videos.

5. REFERENCES

- [1] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, pp. 52–68, March 2011.
- [2] R. Heckel and H. Bolcskei, "Robust subspace clustering via thresholding," *ArXiv e-prints*, July 2013.
- [3] D. Park, C. Caramanis, and S. Sanghavi, "Greedy Subspace Clustering," *ArXiv e-prints*, Oct. 2014.
- [4] O. L. Griffith, B. J. Gao, M. Bilenky, Y. Prichyna, M. Ester, and S. J. M. Jones, "KiWi: A Scalable Subspace Clustering Algorithm for Gene Expression Analysis," *ArXiv e-prints*, Apr. 2009.
- [5] H. Dutta, D. Waltz, K.M. Ramasamy, P. Gross, A. Salieb-Aouissi, H. Diab, M. Pooleery, C.A. Schevon, and R. Emerson, "Patient-specific seizure detection from intra-cranial eeg using high dimensional clustering," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, Dec 2010, pp. 782–787.
- [6] Allen Y. Yang, John Wright, S. Shankar Sastry, and Yi Ma, "Unsupervised segmentation of natural images via lossy data compression," Tech. Rep. UCB/EECS-2006-195, EECS Department, University of California, Berkeley, Dec 2006.
- [7] Yannis Panagakis and Constantine Kotropoulos, "Elastic net subspace clustering applied to pop/rock music structure analysis," *Pattern Recognition Letters*, vol. 38, no. 0, pp. 46–53, 3 2014.
- [8] Wolfgang Hackbusch, *Tensor spaces and numerical tensor calculus*, Springer, 2012.
- [9] Peter D. Hoff, "Separable covariance arrays via the tucker product, with applications to multivariate relational data," *Bayesian Analysis*, vol. 6, no. 2, pp. 179–196, 06 2011.
- [10] Jian Yang, D. Zhang, A.F. Frangi, and Jing-Yu Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.
- [11] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2013/06/30 2009.
- [12] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis of image ensembles," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, June 2003, vol. 2, pp. II–93–9 vol.2.
- [13] Xiaofei He, Deng Cai, and Partha Niyogi, "Tensor subspace analysis," in *NIPS*, 2005.
- [14] Haiping Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Multilinear principal component analysis of tensor objects for recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, vol. 2, pp. 776–779.
- [15] Ulrike Luxburg, "A tutorial on spectral clustering," vol. 17, no. 4, pp. 395–416, 2007.
- [16] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2013.
- [17] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.
- [18] AT&T Laboratories Cambridge, "The Olivetti faces dataset," http://scikit-learn.org/stable/datasets/olivetti_faces.html, 1994.